Cover sheets considered harmful

Emma L. TONKIN [a], Stephanie TAYLOR [a] and Gregory J. L. TOURTE [b]
[a] *UKOLN, The University of Bath*
[b] *The University of Bristol*

Abstract. The spread of the cover sheet is a divisive phenomenon. Their appearance is geographically bound and their content situated in the local political and financial context. In this article we discuss the arguments for and against the coversheet in its guise as a fixture on institutional repository preprints, exploring the issue through statistical information gathered from survey material. We lay out the reasoning behind the use of cover sheets in the United Kingdom and discuss their prevalence and the underlying trends.

Keywords. institutional repositories, cover sheets, survey, copyright, text analysis

Introduction: What are cover sheets?

Cover sheets may appear an unlikely source of controversy. Simply put, they are additional pages added to a resource in a digital repository, almost always prepended to the first page rather than appended as an appendix to the document. The use of cover sheets is often associated with the rise of branding in institutional repositories—that is, they are seen as the solution to the marketing challenge faced by institutional repositories seeking to render the scope and usage patterns of their activities more visible in light of the need to retain funding in a difficult research environment. Institutional repositories, however, face many other challenges, including a [perceived] need to retain a visible copyright statement on each file and to provide user-accessible metadata. In a world in which the majority of resources are accessed directly through links to the resource, rather than through the intermediary of the institutional repository, a cover sheet is the most visible and arguably the most accessible way for institutional repositories to provide useful information to the reader.

Guidance for the institutional repository manager in the United Kingdom has generally been positive on the subject of cover sheets. For example, Brace [3] wrote that the benefits of consistently constructed cover sheets applied either to a subset or a full set of repository objects included the availability of detailed version information (assuming, implicitly, that the cover sheet information set included versioning data) and the possibility of linking to other information such as copyright details. The only downside identified in this instance was the time and resource requirement of the process of cover sheet creation; the solution identified was the use of an automated system for cover sheet creation.

Internationally, viewpoints on cover sheets vary widely. Large archives such as JSTOR provide work behind a cover sheet which provides a standardised view of various information about the document, including citation information. Sites such as Arizona's

DLIST do not embed a full-page cover sheet, although some citation or versioning information may be provided within some documents. The QUT repository in Australia, by comparison, routinely places a cover sheet at the beginning of PDF content. Establishing the prevalence of this practice worldwide would be a major task if completed manually and is perhaps best achieved through reliable, automated analysis.

It is perhaps worth identifying a few uses of cover sheets:

- Content submission cover sheets: for example, grant applications, thesis submissions, even homework submitted by students may be prefaced with a cover sheet identifying the submitter and some basic information about the document. These are a form of process management aid, and inherit from the tradition of the print-out cover sheet, which was widely used in many institutions to identify documents when printed.
- Presentational cover sheets: some types of document, such as reports, may be wrapped or prefaced with an appropriate cover sheet for the purposes of imposing a standard presentation on institutional outputs.
- Cover sheets as an aid for the researcher: having identified and printed or copied the document to the local machine, the researcher has lost potentially useful contextual information about the document. Placing that information within the body of the document means that it is, in principle, retained.
- Cover sheets as a means of author identification: some academic authors feel reassured that their intellectual property and identity as the author of a piece of work is more securely asserted by the use of a cover sheet. In such cases, the cover sheet can be perceived as providing the same author information as a traditional journal article.
- Cover sheets for data papers [9]: A 'data paper' in this context is proposed to be a type of paper that presents 'a neutral, argumentfree platform from which to draw scientific conclusions'. The purpose of a cover sheet in this context is the provision of essential indexing information alongside 'a package of narrative and references that captures the entire data product' (for example, links to the full dataset).
- Cover sheets as a visual reference for copyright and permissions data.
- Cover sheets as a branding exercise.

Cover sheets and copyright

Copyright is a major guiding force behind the use of cover sheets. Matthies and Frisby [10] describe the use of cover sheets as an integral part of the permissions process. A poster presentation by St Andrews [2] regarding the provision of an electronic theses service describes the use of cover sheets not only for branding of the full-text documents, but also to provide relevant context ('anchoring' the full text) and to give reassurance to postgraduates that their work is adequately protected against unauthorised reuse or plagiarism. In other words, there is a perception that cover sheets are either required or beneficial as a visible indicator that a copyright triage/assessment process has been undergone and that any readers will be informed of the copyright status of the work.

In the UK, the SHERPA/RoMEO project [6] provides a series of conditions laid down by publishers for self-archiving. In many of these cases conditions for author self-archiving may include the use of a statement of some kind (i.e., published source must

be acknowledged, a DOI must be provided, etc.). Publishers do not typically specify that this information is to be provided in a cover sheet embedded into the document itself.

What's wrong with cover sheets?

With the exception of cases in which specialised templates are used (such as certain formal report delivery processes, in some cases thesis submissions, and so forth), the use of cover sheets in institutional repositories is almost solely associated with files encoded in fixed-layout formats, particularly PDF. This is probably due to the relative ease and reliability by which these formats can be manipulated on the command line using tools such as pdftk and pdftools. This manipulation is often visible, since cover sheets frequently use a different paper format/layout to the item itself (e.g., letter format vs. A4). The formatting discontinuity is visible to the user and to the printer, and is wholly unjustifiable since there is no reason why, in the cases of common paper formats, a corresponding cover sheet format cannot be generated.

The Version Identification Framework [14] lists a number of pros and cons of cover sheets from the perspective of an enthusiast. The pros listed included

Uniformity The object continues to be identified clearly, including its version status, even if it is removed from the repository itself',
Detail all of the types of essential versioning information [can] be used'
Linking to other repository information, such as policies, that might be related'

The downsides identified included the time and resource commitment required to add cover sheets, which they propose should be offset with automated alternatives, perception of interference (introduction of potentially unwelcome branding, relegation of content to second page), problems with page numbering and, finally, preservation issues. These issues are covered in the following statement: 'Some may view the object as representative of how the research process stands at the time of deposit, and that altering it, even in such a 'behind the scenes' way, is a threat to the integrity of the work in an archiving sense.'

The VIF proposed solution to this last issue is the use of the PREMIS data dictionary [12] for preservation metadata to store information about any changes made. However, the proposed solution requires the identification of what constitutes useful information and what should be retained. For text mining purposes in particular, it is easy for minor changes to significantly alter the usefulness of the file for a given approach.

Consider for example a scenario in which a researcher is attempting to make use of PDFs retrieved from repositories in order to perform author name/identity disambiguation upon a document set: that is, the researcher is seeking to make use of available contextual information in order to tell two people called 'John Smith' apart. Contextual information typically used for this purpose may include location of publication, classification of the text, subject of the text, references contained within the document and so forth. A researcher could equally choose to look at the structure of the document itself and attempt to tell individuals apart by the platform that they use to generate their documents (OpenOffice versus Word; LaTeX on a Mac; printer driver used to print to PDF, Calligra and so forth). It is a simple feature set, easily accessible through manipulation of the PDF format using tools such as pdfinfo or exiftool, but potentially an important one—but this information is usually removed when a cover sheet is generated. As with

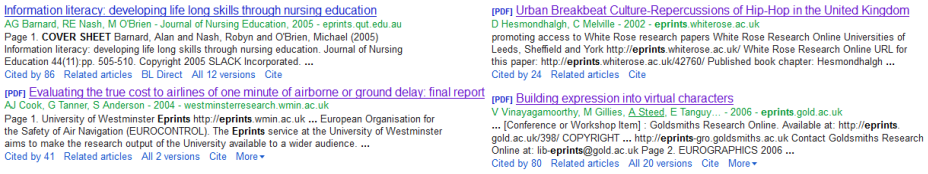

Figure 1. Automated indexing and coversheets

the case of paper size discontinuity, there is no theoretical reason why this information must be lost, since the tools typically used for this purpose are capable of preserving and manipulating metadata.

Consider a second researcher who is looking to index all the documents that he/she has available using an automated indexing tool. The PDF cover sheet presents some difficulties in this context too, since indexing typically picks up the features of the cover sheet rather than the document itself, which may or may not contain the correct information formatted as the indexer expects. Indirectly, the effects of this can be seen on Google Scholar today, since the service makes use of automated indexing as a potential data source (see [5]). The result is that the indexing picks up information from the cover sheet and may, depending on the query, display it in place of more relevant information from the structure of the paper (see fig. 1). In an earlier study, Hewson and Tonkin [7] identified cover sheets as a potential problem for automated indexing and retrieval.

In most cases, Google Scholar is smart enough to resolve this issue satisfactorily. However, the recent spate of articles discussing the low discoverability of articles on institutional repositories [11, 8, 1] raises the question of factors influencing discoverability of articles in repositories; it is perhaps reasonable to wonder whether cover sheets have a positive, negative or neutral impact on the discoverability of items in that repository.

1. Usage of cover sheets: a survey

The rest of this paper is given over to establishing opinions of and usage patterns relating to cover sheets. There are several viable approaches towards this goal. Some are technical, such as the (complete or partial) spidering of institutional repositories for files which can be parsed to identify cover sheets. Initially, we chose to request information directly from repository and CRIS/RIM system managers by means of a questionnaire-based approach. Our primary goals were to elicit the primary motivations for the use of cover sheets, the prevalence of cover sheets in the (predominantly UK) repository landscape, and to identify the purposes for which cover sheets are currently applied. We felt that this would permit us to approach an evaluation of the effectiveness of the approaches taken to fulfil these requirements, allowing us to explore the question of whether and in which contexts cover sheets are effective or, to use Dijkstra's famous 1968 snowclone, may be considered harmful [4].

1.1. Method

In order to investigate the perspectives of repository managers on the issues and practice surrounding cover sheets, we made use of a quantitative survey method. We designed a

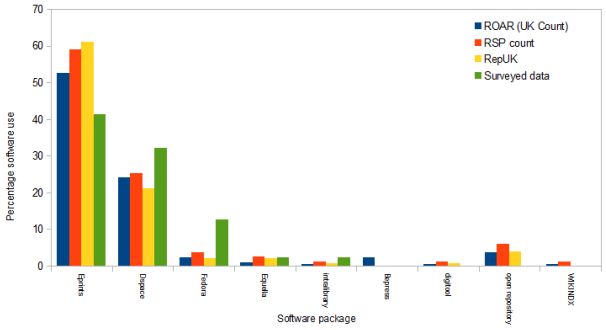

Figure 2. Comparing survey findings with other data sources, such as ROAR, RSP and RepUK

short survey with the intention of learning about digital library and resource management system administrator viewpoints on the subject of the usage, benefits and drawbacks of cover sheets. A qualitative discussion (interview) approach was taken initially with a small subset of repository managers, in order to enable the authors to identify the key issues and concepts to cover during the survey itself, which was initially piloted with a further subset of subjects.

The resulting three-page questionnaire is separated into three sections. Questions 1 and 2 (page 1) explore the repository system used and the use, if any, of cover sheets. The central section discusses managers' views on cover sheets; respondents who answer in the negative to Question 2 complete the survey at this point. Those indicating that they make use of cover sheets instead complete an analogous section discussing the creation and use of cover sheets.

The survey was circulated to three mailing lists with a UK focus, including the JISCMail Repositories list, and was announced on Twitter. It was kept open for a total of four days.

1.2. Results

Overall, 88 respondents filled in the survey. In order to establish the reliability of the survey responses we note firstly that the overall size of the population of UK academic institutional repository managers is proportional to the number of academic repositories. These are variously estimated by the repository indexing service OpenDOAR as summing 209, by ROAR as 250, by RSP as 86 (of which the RSP project [13] publishes a full description of 83) and by RepUK as 150. This variation probably relates to variation in the populations targeted by each service, since RSP includes neither further education colleges nor, geographically, does its coverage include Northern Ireland. The results given here are subject to an approximate ± 7 percent margin of error at the 95% confidence level. Comparing the surveyed data for institutional repository platform usage shows that the survey responses are comparable with other data sources' usage estimates (see Figure 2), with some variation resulting from the survey method, although repeated methods ANOVA shows no statistically significant difference, $F(8,3) = 0.194, \quad p = 0.899$.

Anecdotally, it is suggested that Bepress is more widely used in further education in the UK context, whilst ePrints may attract less 'hacker culture' than Fedora, the inci-

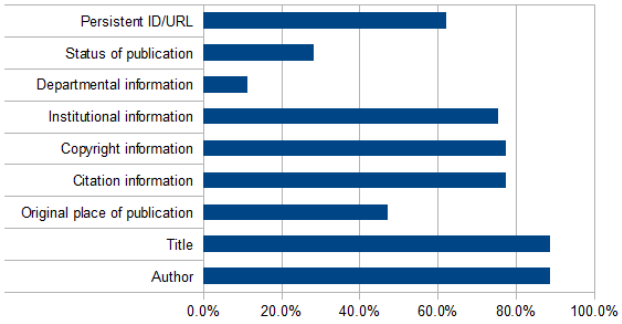

Figure 3. Metadata included within cover sheets.

dence of which is anomalously high in our dataset. Whilst the comparison of these platform usage estimates deserves further discussion in another forum, it is adequate for the purposes of this paper to note that ePrints users in particular are likely to be somewhat underrepresented amongst our respondents.

1.3. Reported usage of cover sheets

57% of respondents responded that their repository included cover sheets on documents. An additional 11% of respondents stated that they made some use of cover sheets, used them for a subset of documents, made use of unconventional cover sheets (for example, placing the coversheet at the back of the document) or were considering making use of cover sheets in future. Thus in aggregate over two thirds of respondents either use, intend to use or instruct repository users to add cover sheets.

1.4. Metadata included within cover sheets

Survey responses to the question 'what information do cover sheets contain in your repository system?' presented a spread of pieces of information of which the most popular are document metadata such as title and author, copyright, citation and institution information, followed by persistent IDs, original venue of publication and the status of the publication (see Figure 3). Cover sheets are overwhelmingly used on categories of full-text documents only, with a few respondents indicating that they were also used on metadata-only records.

1.5. Creation process of coversheets

Both automated and manual processes may coexist in a single repository, depending on the context (i.e., it may sometimes be preferable to generate 'standard' sheets for a collection automatically, whilst in other cases a manual process is necessary). 49% of repository managers reported making use of an automated process for coversheet creation, of which 8% report the usage of a batch processing approach based on available metadata. A manual process is used in 53% of cases.

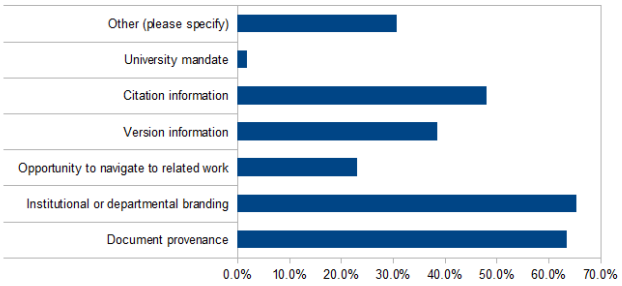

Figure 4. Motivations for usage of cover sheets.

In summary, over half of the repository managers surveyed report that they generate the sheets manually. cover sheets almost always contain author information, title information and copyright information, and the majority of of respondents additionally indicate that citation information is provided, alongside a persistent URL and original place of publication as additional metadata elements provided in cover sheets. In the vast majority of cases, cover sheets are used only on full-text records.

1.6. Motivations for use of cover sheets

Respondents who made use of cover sheets were asked about their primary motivations for doing so (see Figure 4). The most popular reasons were the addition of institutional/ departmental branding and document provenance information, followed by citation information and version information. Although copyright concerns were not provided as an option, around a third of this group identified copyright concerns in a write-in field of the survey, several indicating that the provision of a statement set by the publisher is a requirement for allowing self-archiving or that the provision of a cover sheet is a risk-management strategy.

Respondents who indicated that they do not make use of cover sheets were also surveyed about their views. The majority (over 60%) indicated that they considered cover sheets useful for citation, institutional and copyright indication. Relatively few felt that author information (39%), persistent URL (34%), title, original place of publication (30%) or status of publication (22%) were of use. Around a quarter of these respondents indicated that cover sheets were useful for none of these purposes.

When asked for the primary reasons behind the decision not to implement cover sheets, around a third identified a lack of demand from repository users and/or that the question had not arisen. Other issues mentioned included a lack of support from the repository platform, the scalability of the process/the resources involved, both technical and administrative, low priority and, interestingly, identification of alternative strategies—one respondent noted that they redirect incoming ‘deep’ links to the repository item record, requiring the user to click through the item record page, since this exposes all of the information that would be present on the cover sheet.

Text mining or Google indexing concerns were identified by only one survey respondent, and preservation concerns by none.

2. Discussion

Overall, well over half of all respondents indicated that they make use of cover sheets in at least a subset of cases. The results of our survey clearly demonstrate that there is significant 'buy-in' among repository managers to the concept of the cover sheet as an instrument for compliance with copyright concerns. Indeed, a number of respondents view them as a necessary condition for compliant archiving of preprints. A puzzling aspect of this is the observation that, even in repositories that make extensive use of cover sheets, items in formats that do not permit easy prepending of cover sheets are nonetheless made available, suggesting that they are not a requirement for distribution in at least a subset of cases.

The issues resulting from this practice can usefully be separated into technical and general issues. Of the former, metadata issues relate to loss of evidence and are thus placed in the domain of preservation concerns. Text mining issues relate both to the loss of metadata and potentially document structural data and to the imposed requirement of completing an additional stage of identifying and removing coversheets before proceeding with the task at hand. Layout and formatting issues result in minor inconvenience to the user when viewing and printing the document and are a consequence of incomplete cover sheet generation (i.e., a good cover sheet generation function should be able to detect the format of the original and respond appropriately).

General issues for the repository manager include an ongoing need for technical effort, time/resource requirements, scalability issues and difficulties encountered when upgrading or changing platforms. Repository contributor/user viewpoints on the cover sheet may, as previously indicated, not be straightforward to characterise, as they are viewed as anything from a waste of paper to a useful protective measure; we will not attempt to cover these in more detail here, as they deserve detailed study in their own right.

For a repository manager looking for alternatives to the use of cover sheets, the following options have been explicitly identified by some respondents and through interviews:

- placing relevant information on the repository item page, instead of altering the item itself
- forcing the visitor to click-through the associated repository item page,
- embedding relevant data within the document metadata itself
- use of a smaller and less intrusive repository 'stamp' somewhere on the document, enabling branding and provision of relevant information
- provision of unaltered versions of the documents contained within the repository for access by means other than direct browsing over the Web, permitting those with a use for unaltered versions of the documents to retrieve the originals.

Of these, the third and fourth are as achievable within PDF as the application of cover sheets themselves (and are no more applicable to other file formats such as DOC files). The second resolves part of the issue, ensuring that users view the copyright statement/repository branding, but if that document is distributed elsewhere, it will not contain relevant embedded data.

It is perhaps worth noting that whilst it is broadly understood by the majority of survey respondents that cover sheets often indeed contain useful information, the use of

the cover sheet to protect a document's copyright attribution is not supported by the technical application. Cover sheets may easily be removed in bulk or simply by using 'print to file'; on the technical level they provide significantly less of a challenge to remove than a watermark or a stamp. A visible cover sheet provides only the reassurance that the reader may be assumed to be aware of the information contained within it. That said, the cover sheet seems to have become an agreed convention for fulfilling the compliance requirements of all parties with an interest in rights management. As such, it can be said to have become a 'quick win' way of dealing with these issues for busy IR managers. Establishing and maintaining a good working relationship with authors and rights holders is an essential part of the job for IR managers, so any quick, easy and agreed way of dealing with this potentially tricky area is welcomed into their busy work flow.

3. Conclusions

On the evidence presented here, the adoption of cover sheets is widespread in the UK, although far from uniform. There is a strong argument for the use of some mechanism to present relevant information, as publishers often request that this is done; however, it is not clear that this information must be embedded within each digital object. If this proves to be the case, it will raise questions about the provision of other types of document such as data files in formats that do not support a cover sheet.

In conclusion, the authors recommend a number of possible steps for repository managers. Consider alternative ways to brand your documents, such as using a watermark or a 'stamp' such as that used by ArXiv. Explore the possibility of holding metadata currently contained in the cover sheet elsewhere in the document, such as in metadata fields specific to that file format. Review the performance of your cover sheet generation process, if you use one, and discuss any technical issues identified such as loss of metadata or formatting issues with the repository vendor or support community. Finally ensure that the original versions of files are retained, as you may wish to change your approach in future.

Future work in this area should include a broader survey of cover sheet use worldwide, as well as focusing in some detail on the opinions of repository users and contributors. However, the authors would like to emphasise once again that repository managers don't make the rules; the repository manager is tasked with identifying and applying an appropriate compromise between the concerns of the different stakeholder groups involved, which is not a trivial undertaking.

References

[1] K. Arlitsch and P. S. O'Brien. Invisible institutional repositories: addressing the low indexing ratios of IRs in Google. *Library Hi Tech*, 30(1):60–81, 2012.

[2] J. Aucock. Electronic theses at the university of st andrews: institutional infrastructure, policy and support to establish an electronic theses service. Technical report, Univeristy of St Andrews, 2008.

[3] J. Brace. Versioning in Repositories: Implementing Best Practice. *Ariadne*, 56, 2008. ISSN 1361-3200. URL `http://www.ariadne.ac.uk/issue56/brace`.

[4] E. W. Dijkstra. Letters to the editor: go to statement considered harmful. *Commun. ACM*, 11(3):147–148, March 1968. ISSN 0001-0782. doi: 10.1145/362929.362947.

[5] Google. Indexing in Google Scholar, 2013. URL `http://scholar.google.co.uk/intl/en/scholar/inclusion.html#indexing`.

[6] A. Hanlon and M. Ramirez. Asking for Permission: A Survey of Copyright Workflows for Institutional Repositories. *portal: Libraries and the Academy*, 11(2):683–702, April 2011. URL `http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v011/11.2.hanlon.html`.

[7] A. Hewson and E. Tonkin. Supporting PDF accessibility evaluation: early results from the FixRep project. In *2nd Qualitative and Quantitative Methods in Libraries International Conference (QQML2010)*, 2010.

[8] P. Jacsó. Metadata mega mess in Google Scholar. *Online Information Review*, 34 (1):175–191, 2010. ISSN 1468-4527. doi: 10.1108/14684521011024191.

[9] J. A. Kunze, P. Cruse, R. Hu, S. Abrams, K. Hastings, C. Mitchell, and L. R. Schiff. Practices, Trends, and Recommendations in Technical Appendix Usage for Selected Data-Intensive Disciplines. Technical report, University of California, 2011. URL `http://escholarship.org/uc/item/9jw4964t`.

[10] B. Matthies and K. Frisby. Creating an Institutional Repository "on the Cheap". In *Ohio Valley Group of Technical Services Librarians 2009 Conference*, May 2009.

[11] M. Norris, C. Oppenheim, and F. Rowland. Finding open access articles using Google, Google Scholar, OAIster and OpenDOAR. *Online Information Review*, 32 (6):709–715, 2008. ISSN 1468-4527. doi: 10.1108/14684520810923881.

[12] PREMIS. PREMIS Data Dictionary for Preservation Metadata, 2011. URL `http://www.loc.gov/standards/premis/`.

[13] Repository Support Project (RSP). Cover sheets, 2011. URL `http://www.rsp.ac.uk/documents/briefing-papers/2011/coversheets_RSP_0811.pdf`.

[14] VIF. Covert sheet, January 2008. URL `http://www2.lse.ac.uk/library/vif/framework/Object/cover.html`.